AI. Lec 1 - Lec 2.

Introduction

1. A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting, what is T?

- ( ) The probability of it correctly predicting a future date's weather.
- ( ) The process of the algorithm examining a large amount of historical weather data.
- ( ) None of these.
- (•) The weather prediction task.

2. The amount of rain that falls in a day is usually measured in either millimeters (mm) or inches. Suppose you use a learning algorithm to predict how much rain will fall tomorrow. Would you treat this as a classification or a regression problem?

- (•) Regression
- ( ) Classification

3.

Suppose you are working on stock market prediction. You would like to predict whether or not a certain company will win a patent infringement lawsuit (by training on data of companies that had to defend against similar lawsuits). Would you treat this as a classification or a regression problem?

- (•) Classification
- ( ) Regression

super	unsuper
(0,1) / قيم مستمرة	
مصنفة labled date	unlabled غير موصوف
give the right answers	don't has the rigit answer

4. Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.

- [x] Have a computer examine an audio clip of a piece of music, and classify whether or not there are vocals (i.e., a human voice singing) in that audio clip, or if it is a clip of only musical instruments (and no vocals). ⇒ آة أو لا

- [x] Given genetic (DNA) data from a person, predict the odds of him/her developing داء السكر diabetes over the next 10 years. عنده سكر او لا

- [ ] Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.

- [ ] Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug, and if so what these categories are.

5. Which of these is a reasonable definition of machine learning?

- ( ) Machine learning means from labeled data.
- ( ) Machine learning is the science of programming computers.
- (•) Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.
- ( ) Machine learning is the field of allowing robots to act intelligently.

A I. lec 3

Linear Regression With One Variable

1. Consider the problem of predicting how well a student does in her second **True** year of college/university, given how well they did in their first year. Specifically, let x be equal to the number of "A" grades (including A-. A and A+ grades) that a student receives in their first year of college (freshmen year). We would like to predict the value of y, which we define as the number of "A" grades they get in their second year (sophomore year). Questions 1 through 4 will use the following training set of a small sample of different students' performances. Here each row is one training example. Recall that in linear regression, our hypothesis is

$h_\theta(x)=\theta_0+\theta_1 x$, and we use m
to denote the number of training examples.

	x	y
1	5	4
2	3	4
3	0	1
4	4	3

For the training set given above, what is the value of m? In the box below, please enter your answer (which should be a number between 0 and 10).

4

2. Many substances that can burn (such as gasoline and alcohol) have a chemical structure based on carbon atoms; for this reason they are called hydrocarbons. A chemist wants to understand how the number of carbon atoms in a molecule affects how much energy is released when that molecule combusts (meaning that it is burned). The chemists obtains the dataset below. In the column on the right, "kJ/mol" is the unit measuring the amount of energy released. **True**

$\boxed{h = \theta_0 + \theta_1 x}$

Name of molecule	Number of hydrocarbons in molecule (x)	Heat release when burned (kJ/mol) (y)
methane	1	-890
ethene	2	-1411
ethane	2	-1560
propane	3	-2220
cyclopropane	3	-2091
butane	4	-2878
pentane	5	-3537
benzene	6	-3268
cycloexane	6	-3920
hexane	6	-4163
octane	8	-5471
napthalene	10	-5157

You would like to use linear regression ($h_\theta(x)=\theta_0+\theta_1 x$) to estimate the amount of energy released (y) as a function of the number of carbon atoms (x). Which of the following do you think will be the values you obtain for θ_0 and θ_1? You should be able to select the right answer without actually implementing linear regression.

1	2	3		
~~-38.7~~	~~492.2~~	1023.1	$\theta_0=-569.6, \theta_1=530.9$	-38.7 -569.6 + 530.9 =
-1249.1	~~-718.2~~	-1875	$\theta_0=-1780.0, \theta_1=530.9$	2310.9
~~-2310.9~~	-2841.8	-3372.7	$\theta_0=-1780.0, \theta_1=-530.9$	1749.1
-1100.5	-1631.4	-2162.3	$\theta_0=-569.6, \theta_1=-530.9$	-990.5
-890	-1411	-1560 y		

$$\frac{1}{2m} \sum (h-y)^2$$

باخذ أقرب قيمة تقريبية

$\boxed{-890}$ y لـ h يمكن ان تساوي ال

3. Suppose we set $\theta_0=-1, \theta_1=0.5$. What is $h_\theta(4)$? **Ture** $h_\theta(x) = \theta_0 + \theta_1 X$

$h_\theta(4) = -1 + 0.5 * 4 = 1$

1

$\theta_0 + \theta_1 X = -1 + 0.5 * 4 = 1$

4. Let f be some function so that

$f(\theta_0,\theta_1)$ outputs a number. For this problem,
f is some arbitrary/unknown smooth function (not necessarily the cost function of linear regression, so f may have local optima). Suppose we use gradient descent to try to minimize $f(\theta_0,\theta_1)$ as a function of θ_0 and θ_1. Which of the following statements are true? (Check all that apply.) **True**

- [x] If the learning rate α is too small, then gradient descent may take a very long time to converge. ✓

- [ ] Even if the learning rate α is very large, every iteration of gradient descent will decrease the value of $f(\theta_0,\theta_1)$. increse X

- [ ] If θ_0 and θ_1 are initialized so that $\theta_0=\theta_1$, then by symmetry (because we do simultaneous updates to the two parameters), after one iteration of gradient descent, we will still have $\theta_0=\theta_1$. X

- [x] If θ_0 and θ_1 are initialized at a local minimum, then one iteration will not change their values. ✓ ⇒ لان التفاضل هيساوي الـ Zero

True

- [x] If θ_0 and θ_1 are initialized at the global minimum, then one iteration will not change their values. ✓

- [ ] No matter how θ_0 and θ_1 are initialized, so long as α is sufficiently small, we can safely expect gradient descent to converge to the same solution. X

large	small
increse	decrease.

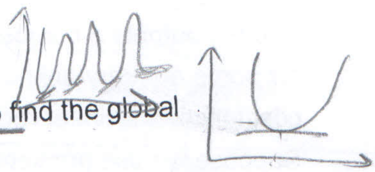
- [x] If the first few iterations of gradient descent cause $f(\theta_0,\theta_1)$ to **increase** rather than decrease, then the most likely cause is that we have set the learning rate α to too large a value. ✓

☐ Setting the learning rate α to be very ~~small~~ large is not harmful, and can only speed up the convergence of gradient descent ✗

يسرع

5. Suppose that for some linear regression problem (say, predicting housing prices as in the lecture), we $(h-y)^2$

have some training set, and for our training set we managed to find some θ_0, θ_1 such that $J(\theta_0,\theta_1)=0$. Which of the statements below must then be true? (Check all that apply.)**true**



☑ Gradient descent is likely to get $\boxed{\text{stuck}}$ at a local minimum and fail to find the global minimum. ✓

☐ Our training set can be fit perfectly by a straight line, i.e., all of our training examples lie perfectly on some straight line. ✗ ⇒ ممكن يأخذ شكل منحنى او داله ٣ او داله او ٤ .

☐ For this to be true, we must have $\boxed{\theta_0=0 \text{ and } \theta_1=0}$ → ليس مهم so that $h_\theta(x)=0$

الثقافة مثل هوال zero ...

☐ For this to be true, we must have $y_{(i)}=0$ for every value of $i=1,2,\ldots,m$. ✗

☐ This is not possible: By the definition of $J(\theta_0,\theta_1)$, it is not possible for there to exist θ_0 and θ_1 so that $J(\theta_0,\theta_1)=0$ ✗

☑ For these values of θ_0 and θ_1 that satisfy $J(\theta_0,\theta_1)=0$, we have that $\boxed{h_\theta(x_{(i)})=y_{(i)}}$ for every training example $(x_{(i)},y_{(i)})$ $J=\frac{1}{2m}\sum(h-y)^2$ ✓

☐ We can perfectly predict the value of y even for new examples that we have not yet seen. ✗
(e.g., we can perfectly predict prices of even new houses that we have not yet seen.)

☐ For this to be true, we must have $\boxed{\theta_0=0 \text{ and } \theta_1=0}$ so that $h_\theta(x)=0$ ✗

$(h-y) \to 0$

Al. Lec 4 -

Linear Regression With Multiple Variables

1. Suppose $m=4$ students have taken some class, and the class had a midterm exam and a final exam. You have collected a dataset of their scores on the two exams, which is as follows:

X_1 midterm exam	$(X_1)^2$ (midterm exam)2	Y final exam
89	7921	96
72	5184	74
94	8836	87
69	4761	78

You'd like to use polynomial regression to predict a student's final exam score from their midterm exam score. Concretely, suppose you want to fit a model of the form $h_\theta(x)=\theta_0+\theta_1x_1+\theta_2x_2$, where x_1 is the midterm score and x_2 is (midterm score)2. Further, you plan to use both feature scaling (dividing by the $\boxed{\text{"max-min"}}$, or range, of a feature) and mean normalization. $X_1(3)$

What is the normalized feature $x(3)1$? (Hint: midterm = 89, final = 96 is training example 1.) Please enter your answer in the text box below. If applicable, please provide at least two digits after the decimal place.

$$\frac{\text{actual - average mean}}{\text{max - min}} \Rightarrow \frac{\boxed{94} - \frac{89+72+94+69}{4}}{94-69}$$

0.52

2. You run gradient descent for 15 iterations with $\boxed{\alpha=0.3}$ and compute $J(\theta)$ after each iteration. You find that the value of $J(\theta)$ **increases** over time. Based on this, which of the following conclusions seems most plausible?

- (•) Rather than use the current value of α, it'd be more promising to try a smaller value of α (say $\alpha=0.1$).
- ( ) Rather than use the current value of α, it'd be more promising to try a larger value of α (say $\alpha=1.0$).
- ( ) $\alpha=0.3$ is an effective choice of learning rate.

3. Suppose you have m=28 training examples with n=4 features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is $\theta=(X^TX)^{-1}X^Ty$. For the given values of m and n, what are the dimensions of θ, X, and $y$in this equation?

- ○ X is 28×5, y is 28×5, θ is 5×5
- ○ X is 28×4, y is 28×1, θ is 4×4
- ◉ X is 28×5, y is 28×1, θ is 5×1
- ○ X is 28×4, y is 28×1, θ is4×1

$X = 28 * 5$

$y = 28 * 1$

$\theta = [\underbrace{(5*28)(28*5)}_{5*5}]^{-1}[(5*28)*(28*1)] = 5*1.$

4. Suppose you have a dataset with m=1000000 examples and n=200000 features for each example. You want to use multivariate linear regression to fit the parameters θ to our data. Should you prefer gradient descent or the normal equation?

- × ○ The normal equation, since gradient descent might be unable to find the optimal θ.
- ◉ Gradient descent, since $(X^TX)^{-1}$ will be very slow to compute in the normal equation. m, n large too slow
- ○ Gradient descent, since it will always converge to the optimal θ. ✓ ×
- ○ The normal equation, since it provides an efficient way to directly find the solution.

$m = 50, n = 10 \Rightarrow$ o o o o

5. Which of the following are reasons for using feature scaling?

- [ ] It speeds up gradient descent by making each iteration of gradient descent less expensive to compute.
- [x] It speeds up gradient descent by making it require fewer iterations to get to a good solution. اوصل لاقرب حل
- [ ] It is necessary to prevent the normal equation from getting stuck in local optima.
- [ ] It prevents the matrix X^TX (used in the normal equation) from being non-invertable (singular/degenerate).

$-1+0.5*4=$

2. Consider the following training set of $m=4$ training examples: **True**

x	y
1	0.5
2	1
4	2
0	0

Consider the linear regression model $h_\theta(x)=\theta_0+\theta_1 x$. What are the values of θ_0 and θ_1 that you would expect to obtain upon running gradient descent on this model? (Linear regression will be able to fit this data perfectly.)

$X=\boxed{1}$ $X=\boxed{2}$

- $\theta_0=1,\theta_1=1$ → 2
- $\theta_0=1,\theta_1=0.5$ → 1.5
- $\theta_0=0.5,\theta_1=0$ → $\boxed{0.5}$ → 0.5
- $\theta_0=0,\theta_1=0.5$ → $\boxed{0.5}$ → $\boxed{1}$
- $\theta_0=0.5,\theta_1=0.5$ → 1

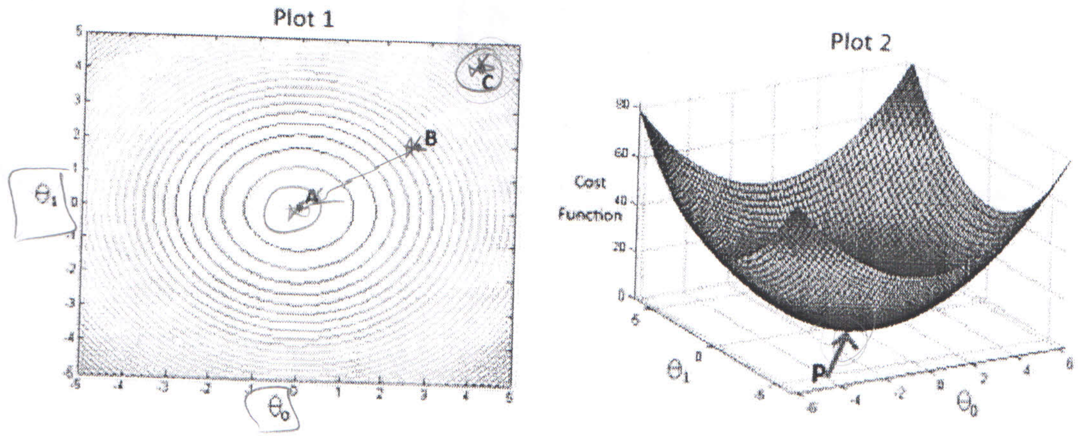
3. Suppose we set $\theta_0=-1,\theta_1=2$. What is $h_\theta(6)$? **True**

11

$\theta_0+\theta_1 X = -1 + 2*6 = 11$

4. In the given figure, the cost function $J(\theta_0,\theta_1)$ has been plotted against θ_0 and θ_1, as shown in 'Plot 2'. The contour plot for the same cost function is given in 'Plot 1'. Based on the figure, choose the correct options (check all that apply). **True**

Plots for Cost Function $J(\theta_0, \theta_1)$



☑ If we start from point B, gradient descent with a well-chosen learning rate (α) will eventually help us reach at or near point A, as the value of cost function $J(\theta_0,\theta_1)$ is <u>minimum</u> at A. ✓

☐ Point P (The global minimum of plot 2) corresponds to point ~~C~~ A of Plot 1. ✗

☐ If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point A, as the value of cost function $J(\theta_0,\theta_1)$ is <u>maximum</u> at point A. ✗
minimum

☐ If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point ~~C~~ A, as the value of cost function $J(\theta_0,\theta_1)$ is minimum at point C. ✗

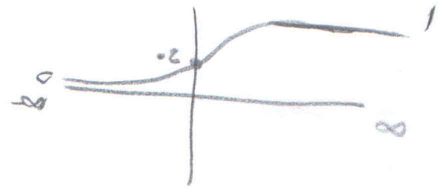
☑ Point P (the global minimum of plot 2) corresponds to point A of Plot 1. ✓

AI.lec5 - lec6

Logistic Regression

1. Suppose that you have trained a logistic regression classifier, and it outputs on a new example x a prediction $h_\theta(x) = 0.2$. This means (check all that apply):

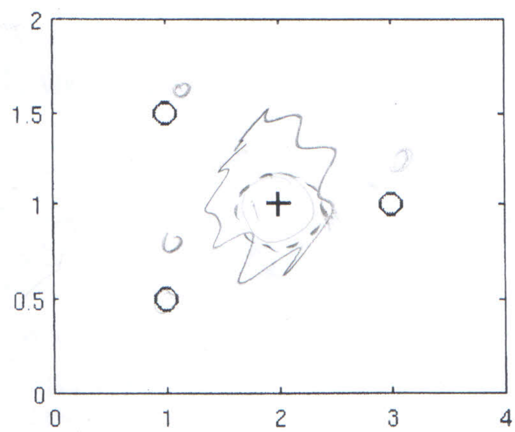
- [x] Our estimate for $P(y=1|x;\theta)$ is 0.2.
- [ ] Our estimate for $P(y=0|x;\theta)$ is 0.2.
- [x] Our estimate for $P(y=0|x;\theta)$ is 0.8.
- [ ] Our estimate for $P(y=1|x;\theta)$ is 0.8.



2. Suppose you have the following training set, and fit a logistic regression classifier $h_\theta(x)=g(\theta_0+\theta_1x_1+\theta_2x_2)$.

x_1	x_2	y
1	0.5	0
1	1.5	0
2	1	1
3	1	0

Which of the following are true? Check all that apply.



- [x] Adding polynomial features (e.g., instead using $h_\theta(x)=g(\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_1^2+\theta_4x_1x_2+\theta_5x_2^2)$) could increase how well we can fit the training data.**true** ⇒ دائرة

$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \boxed{\theta_3 x_1^2} + \theta_4 x_1 x_2 + \boxed{\theta_5 x_2^2}$

logistic

error ≥ 0

- [x] At the optimal value of θ (e.g., found by fminunc), we will have $J(\theta) \geq 0$. **true**
- [ ] Adding polynomial features (e.g., instead using $h_\theta(x)=g(\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_1^2+\theta_4x_1x_2+\theta_5x_2^2)$) would increase $J(\theta)$ because we are now summing over more terms.

$\boxed{0 < h(\theta) < 1}$

- [ ] If we train gradient descent for enough iterations, for some examples $x^{(i)}$ in the training set it is possible to obtain $h_\theta(x^{(i)})>1$. $J(\theta)$ will be a convex function, so gradient descent should converge to the $\boxed{\text{global minimum.}}$ X
- [ ] X The positive and negative examples cannot be separated using a straight line. So, gradient descent will fail to converge.
- [ ] X Because the positive and negative examples cannot be separated using a straight line, linear regression will perform as well as logistic regression on this data.
- [x] $J(\theta)$ will be a convex function, so gradient descent should converge to the (global minimum.)

non-convex ... local min

3. For logistic regression, the gradient is given by $\frac{\partial}{\partial\theta_j}J(\theta)=\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})x_j^{(i)}$. Which of these is a correct gradient descent update for logistic regression with a learning rate of α? Check all that apply.

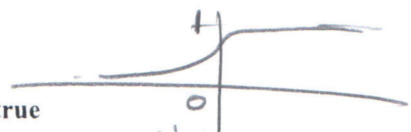
$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$

$\theta = \theta - \alpha \frac{dJ(\theta_0,\theta_1)}{d\theta}$

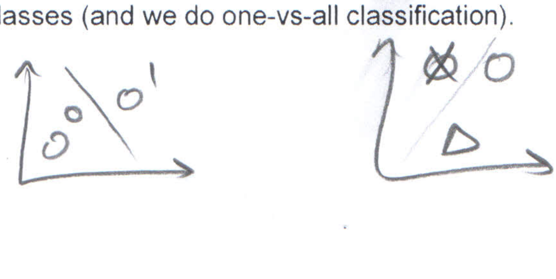
- [ ] X $\theta:=\theta-\alpha\frac{1}{m}\sum_{i=1}^{m}(\theta^Tx-y^{(i)})x^{(i)}$.
- [ ] X $\theta_j:=\theta_j-\alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})x^{(i)}$ (simultaneously update for all j).
- [x] $\theta_j:=\theta_j-\alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})x_j^{(i)}$ (simultaneously update for all j).
- [x] $\theta_j:=\theta_j-\alpha\frac{1}{m}\sum_{i=1}^{m}(\frac{1}{1+e^{-\theta^Tx^{(i)}}}-y^{(i)})x_j^{(i)}$ (simultaneously update for all j).

$\frac{1}{m}\sum(h_\theta(x^{(i)}) - y^{(i)})x(i)$

4. Which of the following statements are true? Check all that apply.



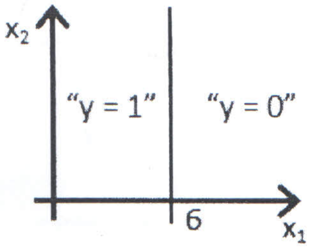
- [x] The sigmoid function $g(z)=\frac{1}{1+e^{-z}}$ is never greater than one (>1). **true**
- [ ] Linear regression always works well for classification if you classify by using a threshold on the prediction made by linear regression. ⇒
- [x] The cost function $J(\theta)$ for logistic regression trained with $m \geq 1$ examples is always greater than or equal to zero. **true**
- [ ] For logistic regression, sometimes gradient descent will converge to a local minimum (and fail to find the global minimum). This is the reason we prefer more advanced optimization algorithms such as fminunc (conjugate gradient/BFGS/L-BFGS/etc). X
- [ ]
- [ ] Since we train ~~one~~ two classifier when there are two classes, we train two classifiers when there are three classes (and we do one-vs-all classification).



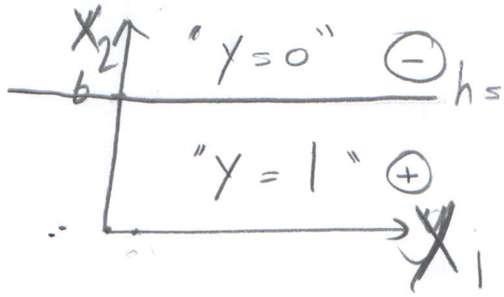
The one-vs-all technique allows you to use logistic regression for problems in which each $y(i)$ comes from a fixed, discrete set of values. **true**

5. Suppose you train a logistic classifier $h_\theta(x)=g(\theta_0+\theta_1 x_1+\theta_2 x_2)$. Suppose $\theta_0=6, \theta_1=0, \theta_2=-1$. Which of the following figures represents the decision boundary found by your classifier?

Figure:



$h(\theta)\ 6 = x_2$



① معنديش x_1 ← فقط
جميع قيم الـ x_2 = 0 وبرسم
المعادلة
$h(\theta) = \boxed{6}$

② يعوض x_2

Figure:

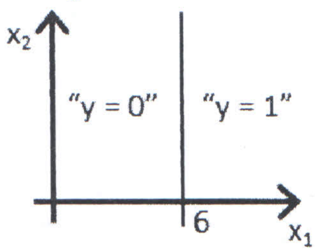
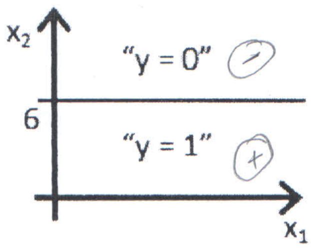
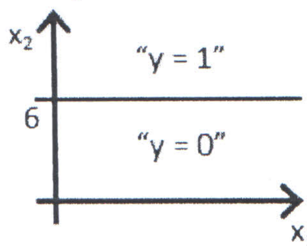


Figure: **true**



$6 - x_2$

Figure:

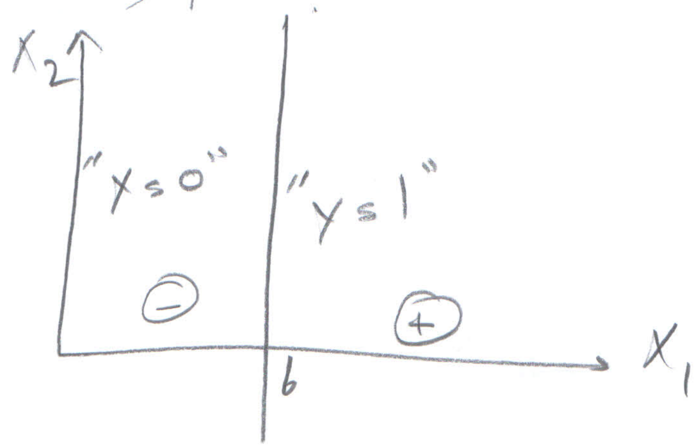


5.

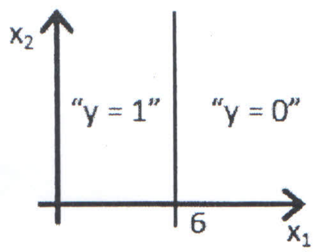
Suppose you train a logistic classifier $h_\theta(x)=g(\theta_0+\theta_1 x_1+\theta_2 x_2)$. Suppose $\theta_0=-6, \theta_1=1, \theta_2=0$. Which of the following figures represents the decision boundary found by your classifier?

$h_\theta(x) = g(-6 + x_1)$

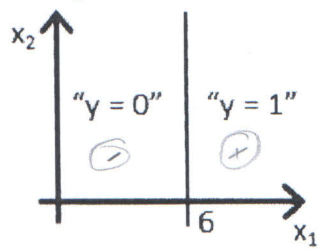
$= \cancel{\frac{1}{1-e^{-\theta T X}}}$



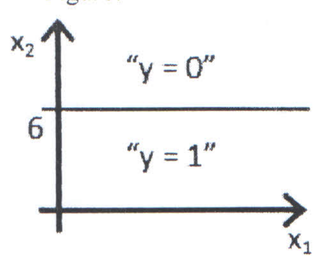
- ○ Figure:



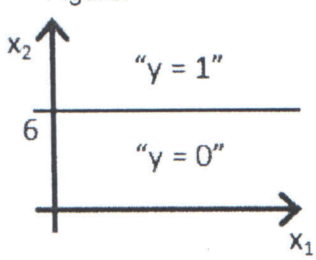
- ◉ Figure:



- ○ Figure:



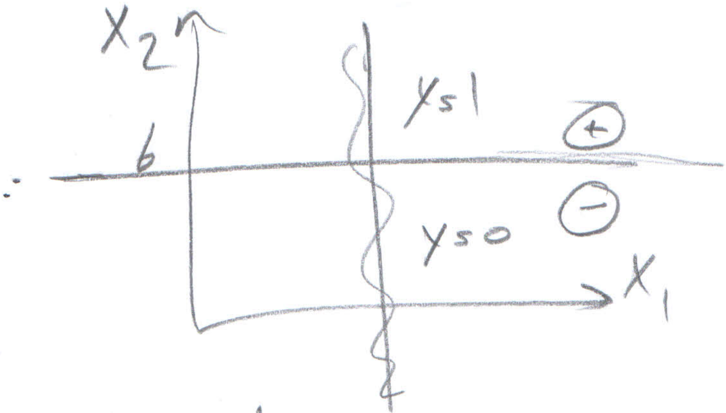
- ○ Figure:



greater than or equal to zero.

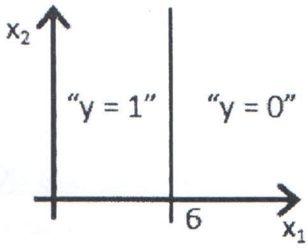
5. Suppose you train a logistic classifier $h_\theta(x)=g(\theta_0+\theta_1 x_1+\theta_2 x_2)$. Suppose $\theta_0=-6, \theta_1=0, \theta_2=1$. Which of the following figures represents the decision boundary found by your classifier?

$h = -6 + x_2$

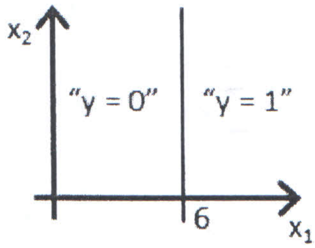


$h = \frac{1}{1+e^{-(6+x_2)}}$

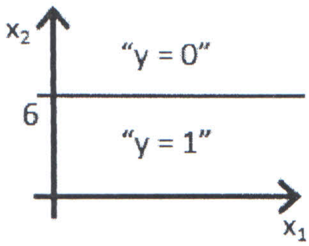
- ( ) Figure:



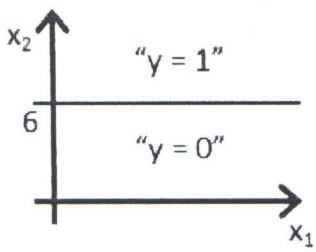
- ( ) Figure:



- ( ) Figure:



- (•) Figure:



AI. lec 6 - lec 7

Regularization

1. You are training a classification model with logistic regression. Which of the following statements are true? Check

all that apply.

☐ Adding a new feature to the model always results in equal or better performance on examples not in the training set.

☐ Introducing regularization to the model always results in equal or better performance on examples not in the training set.

☐ Introducing regularization to the model always results in equal or better performance on the training set.

☑ Adding many new features to the model makes it more likely to overfit the training set.

2. Suppose you ran logistic regression twice, once with $\lambda=0$, and once with $\lambda=1$. One of the times, you got

parameters $\theta=[23.4 \quad 37.9]$, and the other time you got $\theta=[1.03 \quad 0.28]$. However, you forgot which value of λ corresponds to which value of θ. Which one do you think corresponds to $\lambda=1$?

◉ $\theta=[1.03 \quad 0.28]$

○ $\theta=[23.4 \quad 37.9]$

3. Which of the following statements about regularization are true? Check all that apply.

☑ Using too large a value of λ can cause your hypothesis to underfit the data.

☐ Using a very large value of λ cannot hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems.

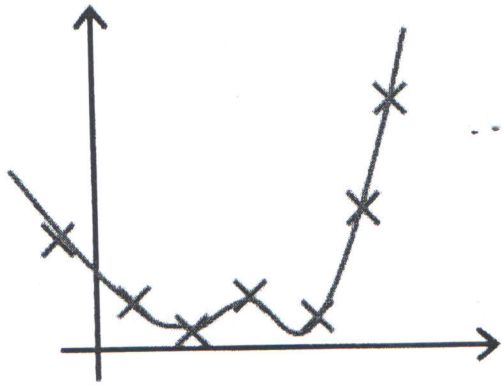
☐ Because regularization causes $J(\theta)$ to no longer be convex, gradient descent may not always converge to the global minimum (when $\lambda>0$, and when using an appropriate learning rate α).

☐ Because logistic regression outputs values $0 \leq h\theta(x) \leq 1$, it's range of output values can only be "shrunk" slightly by regularization anyway, so regularization is generally not helpful for it.

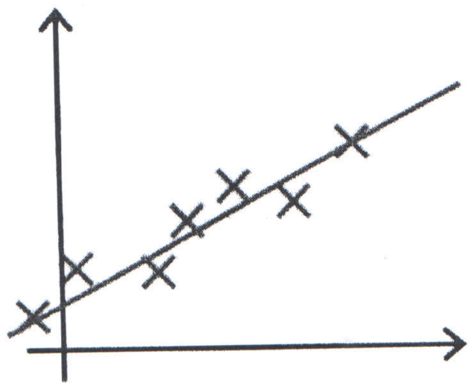
☑ Consider a classification problem. Adding regularization may cause your classifier to incorrectly classify some training examples (which it had correctly classified when not using regularization, i.e. when $\lambda=0$).

4. In which one of the following figures do you think the hypothesis has overfit the training set?

◉ Figure:



○ Figure:



○ Figure:

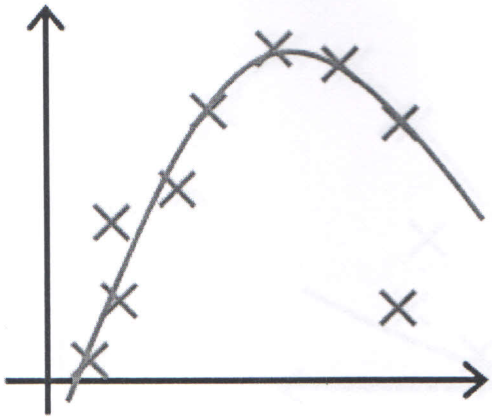
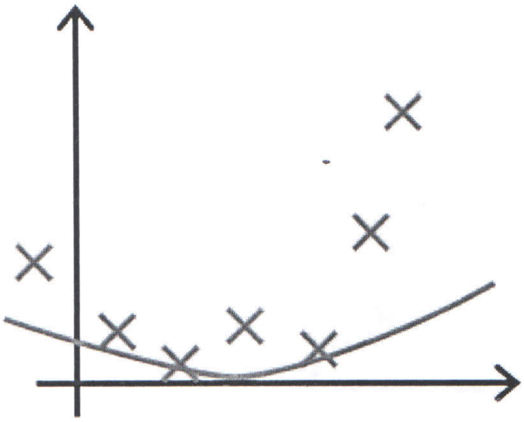


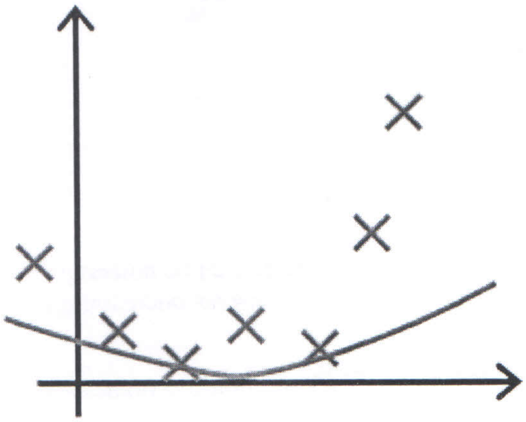
Figure:



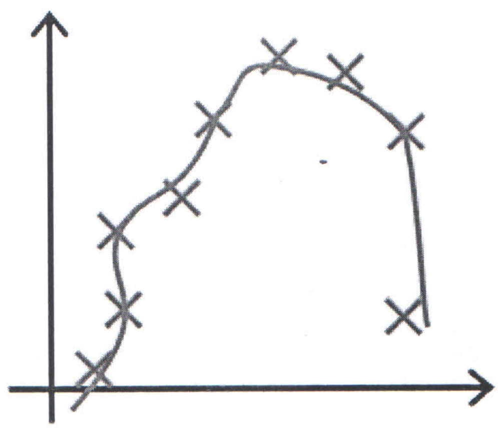
1
point

5.
In which one of the following figures do you think the hypothesis has underfit the training set?

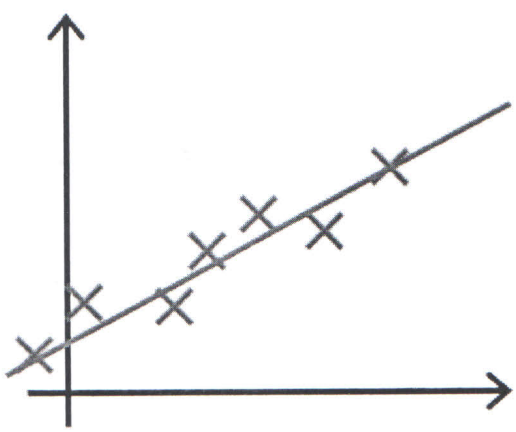
Figure:



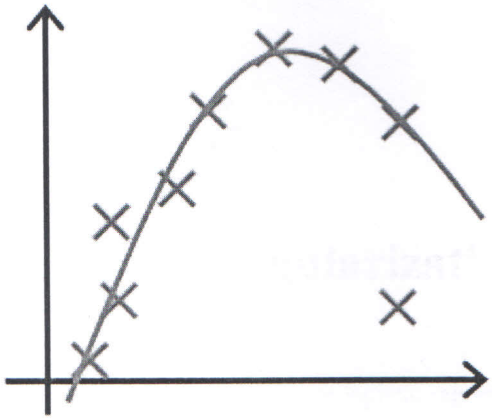
O Figure:



O Figure:



O Figure:



3 questions unanswered

Submit Quiz

Al. lec 8

Unsupervised Learning

1. For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.**true**

- [ ] Given historical weather records, predict if tomorrow's weather will be sunny or rainy.
- [x] From the user usage patterns on a website, figure out what different groups of users exist.
- [ ] Given many emails, you want to determine if they are Spam or Non-Spam emails.
- [x] Given a set of news articles from many different news websites, find out what are the main topics covered.

---

true

- [x] Given a database of information about your users, automatically group them into different market segments.
- [x] Given sales data from a large number of products in a supermarket, figure out which products tend to form coherent groups (say are frequently purchased together) and thus should be put on the same shelf.
- [ ] Given historical weather records, predict the amount of rainfall tomorrow (this would be a real-valued output)
- [ ] Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

True

- [x] From the user usage patterns on a website, figure out what different groups of users exist.
- [x] Given a set of news articles from many different news websites, find out what are the main topics covered.
- [ ] Given historical weather records, predict if tomorrow's weather will be sunny or rainy.
- [ ] Given many emails, you want to determine if they are Spam or Non-Spam emails.

2. Suppose we have three cluster centroids $\mu_1=[1\ \ 2]$, $\mu_2=[-3\ \ 0]$ and $\mu_3=[4\ \ 2]$. Furthermore, we have a training example $x_{(i)}=[1\ \ -2]$. After a cluster assignment step, what will $c_{(i)}$ be?

- ( ) $c_{(i)}$ is not assigned
- ( ) $c_{(i)}=3$
- ( ) $c_{(i)}=2$
- (•) $c_{(i)}=1$

2. Suppose we have three cluster centroids $\mu_1=[1\ \ 2]$, $\mu_2=[-3\ \ 0]$ and $\mu_3=[4\ \ 2]$. Furthermore, we have a training example $x_{(i)}=[3\ \ 1]$. After a cluster assignment step, what will $c_{(i)}$ be?**True**

- ( ) $c_{(i)}$ is not assigned
- (•) $c_{(i)}=3$
- ( ) $c_{(i)}=2$
- ( ) $c_{(i)}=1$

3. K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two? **true**

- [ ] Move each cluster centroid μ_k, by setting it to be equal to the closest training example$x_{(i)}$
- [ ] The cluster centroid assignment step, where each cluster centroid μ_i is assigned (by setting $c_{(i)}$) to the closest training example $x_{(i)}$.
- [x] The cluster assignment step, where the parameters $c_{(i)}$ are updated.
- [x] Move the cluster centroids, where the centroids μ_k are updated.
- [ ] Test on the cross-validation set.
- [x] Randomly initialize the cluster centroids.

4. Suppose you have an unlabeled dataset $\{x_{(1)},\ldots,x_{(m)}\}$. You run K-means with 50 different random initializations, and obtain 50 different clusterings of the

data. What is the recommended way for choosing which one of

these 50 clusterings to use?**True**

- (•) Compute the distortion function $J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_k)$, and pick the one that minimizes this.
- ( ) Plot the data and the cluster centroids, and pick the clustering that gives the most "coherent" cluster centroids.
- ( ) Manually examine the clusterings, and pick the best one.
- ( ) Use the elbow method.

- ( ) The only way to do so is if we also have labels $y^{(i)}$ for our data.
- ( ) Always pick the final (50th) clustering found, since by that time it is more likely to have converged to a good solution.
- (•) For each of the clusterings, compute $\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-\mu_{c^{(i)}}||^2$, and pick the one that minimizes this.
- ( ) The answer is ambiguous, and there is no good way of choosing.

5. Which of the following statements are true? Select all that apply. **true**

- [ ] Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid
- [x] A good way to initialize K-means is to select K (distinct) examples from the training set and set the cluster centroids equal to these selected examples.
- [x] On every iteration of K-means, the cost function $J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_k)$ (the distortion function) should either stay the same or decrease; in particular, it should not increase.
- [ ] K-Means will always give the same results regardless of the initialization of the centroids.

True

- [ ] The standard way of initializing K-means is setting $\mu_1=\dots=\mu_k$ to be equal to a vector of zeros.
- [ ] Since K-Means is an unsupervised learning algorithm, it cannot overfit the data, and thus it is always better to have as large a number of clusters as is computationally feasible.
- [x] For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.
- [x] If we are worried about K-means getting stuck in bad local optima, one way to ameliorate (reduce) this problem is if we try using multiple random initializations.